Monday, June 10, 2019

Judge Janet T. Neff
United States District Judge
Western District of Michigan

Dear Judge Neff,

On March 4, 2019 you ordered that Dr. Mike Coble and I provide to you "a succinct written report and independent opinions on the use of STRmix probabilistic genotyping software in forensic DNA analysis generally," and in the case of US v. Daniel Gissantaner, specifically in regard to five specific sets of topics. I will speak to each of those topics in the five sections of this report below.

I have undertaken the preparation of this report with the understanding that I owe an overriding duty to the Court to provide independent assistance by way of unbiased opinion in relation to the topics on which I have been asked to comment. I have reviewed almost 700 pages of transcripts from a Daubert hearing that took place on May 23 and 24, 2018 as well as what I understand to be a complete set of Defendant's and Government's admitted exhibits.

Section 1. <u>Testing:</u> Whether the use of STRmix has been adequately tested and validated independently of the testing by the developer.

Testing and validation is generally considered to fall into two categories: 1) developmental, and 2) internal. The testing/validating of STRmix by its developer would be considered to be developmental. Two different kinds of internal testing/validation should be of interest to the Court: a) that performed by other laboratories that have chosen to use STRmix in their casework; and b) that performed by the Michigan State Police (MSP) Laboratory prior to its use of STRmix in its casework.

More than 30 laboratories worldwide had performed internal validation studies of STRmix prior to the May, 2018 Daubert hearing for this case. Many had published their findings in a variety of ways including peer-reviewed research journals and meeting presentations/posters. Some of those publications/presentations were independent of the developer though some include individuals such as Dr. Buckleton as a co-author.

Relatively little was said during the May, 2018 Daubert hearing in this case about the MSP Laboratory's testing/validation of STRmix prior to its use in casework beginning in early 2016. But, a 47 page "Validation Summary" for STRmix-PowerPlex Fusion was admitted as Government exhibit 10. I have tried to identify

testing that the MSP Laboratory performed on samples that were similar to the gun swab sample (LS15-377) that was tested in this case[1].

The validation summary suggests that the MSP Laboratory did evaluate a single known three-person mixture where the mixture ratio was 3:2:1 and the lowest level contributor (at 17%) was responsible for either 117 picograms (pg), 78 pg, 58, pg, or 26 pg of template DNA in a series of four experiments (Figure 18). It concluded "It can be noted that under all circumstances, significant likelihood ratios can be obtained with the use of STRmix™ at all DNA amounts tested. However, the significance of the likelihood ratios are negatively impacted as the input DNA amount decreases and the extent of allelic and locus drop-out increases." (page 39) No likelihood ratios appear to have been generated for known non-contributors in these experiments.

The validation summary suggests that the MSP Laboratory also evaluated a single known four-person mixture where the mixture ratio was 4:3:2:1 and the lowest level contributor (at 10%) was responsible for either 117 picograms (pg), 78 pg, 58, pg, or 26 pg of template DNA in a series of four experiments (Figure 19). It again concluded that STRmix generated "significant likelihood ratios regardless of the input DNA amounts." (page 40) No likelihood ratios appear to have been generated for known non-contributors in these experiments.

Other mixtures of three- and four-contributors were evaluated as part of a section titled "Hypothesis testing with contributors and non-contributors" on pages 28 through 38. While some of those mixture ratios would have resulted in the lowest level contributor being responsible for as little as 4% of the template DNA (e.g. a 10:10:5:1 four-person mixture) I am not able to determine from the summary what quantity of DNA they contributed. The number of analyses where the lowest level contributor was responsible for 7% or less of the total amount of template DNA in the MSP Laboratory internal validation study appears to be small – perhaps as few as five to ten. There is no clear indication that in any of those instances did the lowest level contributor add 50 pg or less to the total amount of template DNA that was analyzed.

Dr. Lund testified during the May, 2018 Daubert hearing that the models that are used to describe the relationship between a response (e.g. allelic drop out) and an explanatory variable (e.g. peak height) are only approximations of the real relationship. A scatterplot might look like it is consistent with a linear relationship

---

[1] In her testimony, Amber Smith suggested that the evidence sample contained DNA from three contributors but was open to the possibility that it arose from four contributors. She believed that "around 3 microliters" of a 0.2344 nanogram (ng) per microliter solution was used for testing – corresponding to approximately 0.7 ng of total DNA. STRmix estimated that if the sample did arise from three contributors that the lowest level contributor (the one who might be the person of interest) was responsible for approximately 7% of the total amount of DNA – corresponding to just less than 50 picograms of DNA.

but it would be a mistake to assume that we have information about the nature of the relationship beyond the data collected during a validation study.

Similarly, as pointed out by Dr. Coble et al. in a 2016 publication (Government's exhibit 23), "The goal of an internal validation study is to explore the limitations of the software and test the reliability, robustness, and reproducibility of the system. Samples that mimic the types of cases encountered should be tested." That same Coble et al. 2016 paper goes on to say "Determination of the limits of the software is important to establish the types of profiles that are suitable for handling by the laboratory. . . . Probabilistic software, especially for low-level DNA mixtures, may allow a laboratory to widen the scope of their casework in terms of the type of evidence handled. However, there may also be a temptation to submit all complex mixtures to particularly versatile software. Therefore, the community is reminded of a previous recommendation of the DNA Commission (2) that is still valid: '(Gill et al., 2006, Recommendation 8): If the alleles of certain loci in the DNA profile are at a level that is dominated by background noise, then a biostatistical interpretation for these alleles should not be attempted.'"

The MSP Laboratory validation of STRmix does not appear to have identified any limitations such as at what level alleles are dominated by background noise (at least in terms of quantity of template DNA or mixture ratio) for 2-, 3- and 4-person low-level DNA mixtures.

Section 2. <u>Peer Review:</u> The nature and extent of independent peer review of the use of STRmix.

As was noted during the Daubert hearing in May, 2018, Dr. Buckleton and his colleagues are prolific publishers of peer-reviewed papers. Dr. Buckleton is clearly proud of STRmix and eager to find opportunities to tell others about it.

However, a fairly common criticism of STRmix (as well as of its largest competitors for market share within the US) is that there have been virtually no publications by individuals who were either not directly affiliated with the development and sale of STRmix and/or who have made a financial commitment to utilize STRmix. This undesirable circumstance is likely a by-product of: 1) concerns that programs like STRmix will be misused by those who have not been trained by their proponents, 2) the investment associated with purchasing a license to use STRmix, and 3) restrictions on how STRmix can be used during trial periods.

There is unquestionably a value associated with extensive independent peer review of any powerful analytical tool or approach. It is unlikely that that value will be realized if courts or government procurement guidelines do not insist on truly independent review of both the performance and the coding/implementation of programs like STRmix.

Section 3. <u>Error Rate and Standards:</u>

a) Evidence of the rate of error in applying STRmix and significance to the DNA testing in this case

There was testimony in the May, 2018 Daubert hearing regarding if the likelihood ratio reported for the defendant in this case might be 49 million or as little as 5 million. Reasonable concerns were also discussed about the ease with which errors could be introduced through things such as the Prosecutor's Fallacy.

These concerns could be mitigated by a careful explanation of the values in both the numerator and the denominator of likelihood ratios. Both the numerator and the denominator are estimates of how consistent the test results that were obtained are with what would be expected if the prosecution or the defense hypothesis, respectively, was correct when both are evaluated by the same model (here, the algorithm incorporated into the version of STRmix that was used). If the results are perfectly consistent with the prosecution's theory of the case (e.g. that all of a specific individual's alleles are detected and no detected alleles could have not been contributed by that individual) then the numerator is "1." The converse corresponds to "0." But, the results obtained for evidence samples are not always perfectly consistent with the prosecution's theory of a case (e.g. drop out needs to be invoked to explain the failure to detect some of the defendant's alleles) in which case the numerator is a value less than 1. The STRmix model might determine that the observed results are very inconsistent with what would be expected if the prosecution's theory of the case is correct – but, that same model might simultaneously find that the observed results are on the order of 50 million times more inconsistent with what would be expected if the defense theory of the case is correct. Different models might yield different numerators and/or denominators and all models are just approximations.

There was relatively little testimony about a false positive or false negative error rate per se for STRmix in the transcript of the May 2018 Daubert hearing for this case. Publications included in the Government's exhibits do suggest that fewer than 1 in 1,000 known non-contributors to a sample would be associated with a likelihood ratio of 1,000 or more. By extension, many millions of known non-contributors might need to be evaluated before one was found to have a likelihood ratio greater than the one reported for the defendant in this case.

b) Maintenance of standards, certification and extent of validation of STRmix by the Michigan State Police Laboratory

Please see my remarks in the "Testing" section above in regard to the internal validation of STRmix performed by the MSP Laboratory. The evidence sample in this case appears to fall outside of (below) the ranges of %-contribution and quantity-of-template-contributed for which the MSP Laboratory has validated STRmix.

c) Whether the validation was reviewed by any external auditors

I have not seen an affirmative indication that the MSP Laboratory's validation was reviewed by any external auditors. But, given that their protocols provide for STRmix analyses it would be expected that external auditors/accreditors would have asked to see that the laboratory had performed a validation study.

and d) Is it an accepted protocol that self/internal validation of DNA analysis software by a laboratory is sufficient or does scientific protocol require external and independent review of the validation?

It is common practice for testing laboratories in the US to begin using a new approach/tool for casework after completing an internal validation and often in dozens or even hundreds of cases before an external and independent review of the validation. In my experience, most external and independent reviews of validation studies are performed during admissibility challenges. In such circumstances, those external and independent reviews are often redundant and vary in terms of their rigor and scope.

Section 4. General Acceptance in the Scientific Community

a) What is the relevant scientific community that determines the general acceptance of STRmix

There is a range of opinion on this issue.

On one extreme, it has been argued that only a very small number of individuals who have been intimately involved with the development of probabilistic genotyping software should be considered the relevant scientific community. This is a paradoxical suggestion because most if not all of those individuals require that those who would perform an evaluation sign a non-disclosure agreement that would prevent them from being involved in the development of probabilistic genotyping software.

The 2018 letter to the editor of the Journal of Forensic Sciences (defense exhibit E) for which both Nathan Adams and I are co-signers observes that software-based probabilistic genotyping approaches (like STRmix) "are necessarily rooted in collaboration between experts in the area of molecular biology, population genetics, statistics, forensic science, computer science, and software engineering. While it is important to consider the perspective of all of these disciplines on the validation issue, we think that the perspective of software engineers are particularly important. Decades of experience with software failures have led to established practices for what is commonly known as verification and validation (V&V) of software. We urge that those practices be followed when evaluating PG systems." As such, it is my opinion that experts in all those disciplines (molecular biology, population genetics, statistics, forensic science, computer science, and software

engineering) be considered part of the relevant scientific community that determines the general acceptance of a computer program like STRmix.

and b) What evidence exists of general acceptance of STRmix in that community?

STRmix's use in casework by more than 30 US crime laboratories (including the FBI) at the time of the May, 2018 Daubert hearing for this case is in itself evidence that STRmix has been generally accepted as a viable solution to a difficult problem faced by the forensic science community (specifically, attaching reliable statistical weights to mixed DNA samples with an unknown number of contributors where allelic drop out may have occurred). Proponents of STRmix also point to several successes in admissibility hearings held by a number of state courts.

I agree with Dr. Buckleton's testimony that, of all the probabilistic genotyping systems being used by crime laboratories in the US for casework at this time, STRmix comes the closest to following the IEEE V&V standards that are embraced by software engineering professionals. But, no one is suggesting that STRmix has adhered to IEEE V&V standards. It would be difficult to find a software engineer who maintained that a program whose output has such potential to lead to loss of liberty or life be deemed acceptable simply because alternatives were less rigorously developed. While there are credible alternatives to IEEE's V&V standards, adherence to software development/testing practices that are generally accepted y software-engineering professionals should not be considered a matter of stylistic preference.

Section 5. Application of STRmix in this Case: what are the implications for the use of STRmix probabilistic genotyping software in the circumstances of this case?

In section 3. b. of this report I say "The evidence sample in this case seems to fall outside of (below) the ranges of %-contribution and quantity-of-template-contributed for which the MSP Laboratory has validated STRmix." From the materials that have been provided for me to review it appears that the evidence sample (LS15-377, a gun swab) is a mixture of DNA from at least three individuals where the individual who contributed the least material is responsible for only 7% of the total DNA that was used for testing (approximately 49 pg). These values are well below the levels at which the 2016 PCAST report felt that some probabilistic genotyping systems had been foundationally validated. 49 pg corresponds to about what would be obtained from eight or nine human cells and is much less than what the test kit used by the MSP Laboratory recommends for optimum results. It would be inappropriate to assume that an approach or tool worked reliably outside of the range of values upon which it had been tested.

No one wants to be on an airplane where only some of the software has been developed in accordance with IEEE V&V standards even if the only alternative is an airplane where even less of the software is consistent with those standards. And,

unlike airplanes and automobiles, as is said in the concluding paragraph of defense exhibit E, "software defects in probabilistic genotyping systems will be difficult to detect particularly if critical errors occur only under some circumstances but not others. A problem of this sort might persist without being noticed for a considerable period of time, while doing incalculable damage to the public interest and, ultimately, to the reputation of forensic science."

I look forward to an opportunity to answer the Court's questions directly during a hearing on Monday, July 8, 2019. Please let me know if there is any other way I can be of help.

Sincerely,

Dan E. Krane 6/10/19

Dan E. Krane,
Professor of Biological Sciences, Wright State University
President/CEO, Forensic Bioinformatics